VIB2SOUND: SEPARATION OF MULTIMODAL SOUND SOURCES

Mai Akahoshi[1], *Yuhang Wang*[1,2], *Longbiao Cheng*[1], *Anja T. Zai*[1,2], *Richard H. R. Hahnloser*[1,2]

[1]Institute of Neuroinformatics, University of Zurich and ETH Zurich, 8057 Zurich, Switzerland.
[2]Neuroscience Center Zurich (ZNZ), University of Zurich and ETH Zurich, 8057 Zurich, Switzerland.

ABSTRACT

Understanding animal social behaviors, including vocal communication, requires longitudinal observation of interacting individuals. However, isolating individual-level vocalizations in complex settings is challenging due to background noise and frequent overlaps of coincident signals from multiple vocalizers. A promising solution lies in multimodal recordings that combine traditional microphones with animal-borne sensors, such as accelerometers and directional microphones. These sensors, however, are constrained by strict limits on weight, size, and power consumption and often lead to noisy or unstable signals. In this work, we introduce a neural network-based system for sound source separation which leverages multi-channel microphone recordings and body-mounted accelerometer signals. Using a multimodal dataset of zebra finches recorded in a social setting, we demonstrate that our method outperforms conventional microphone array-based systems. By enabling the separation of overlapping vocalizations, our approach offers a valuable tool for studying animal communication in complex naturalistic environments.

***Index Terms*—** Sound Source Separation, Bioacoustics, Animal Communication, Multimodal Signal Processing

1. INTRODUCTION

Understanding social behaviors in animals requires longitudinal observation of interactions within group settings. Among these behaviors, vocal communication plays a critical role and has attracted widespread interest across disciplines. Traditionally, vocal behaviors are recorded using microphones that capture all sounds generated in the scene. However, in social settings where multiple animals vocalize simultaneously, these recordings often contain overlapping vocalizations and environmental background noise, making it difficult to attribute vocalizations to specific individuals.

Microphone array-based approaches have been proposed to meet this challenge, as microphone arrays capture sound at multiple spatial locations, allowing the exploitation of time and amplitude differences between microphones to estimate the direction or position of a vocal source [1]. Microphone arrays combined with neural networks have shown superior performance in sound source separation tasks in human speech [2, 3], while their application in bioacoustics is relatively limited. Overlapping vocalizations among individuals of the same species are especially hard to separate due to their high acoustic similarity, forcing researchers to exclude such overlapping vocalizations from analysis [4, 5]. To address this issue, body-mounted sensors, such as miniature microphones and accelerometers, have been proposed for use in rats [6], mice [7], birds [8], and humans [9]. Although such sensors can provide valuable individual-level data, they are often compromised in signal quality compared to microphone signals due to strict constraints on weight, size, and power consumption. Also, such sensors often signal low-frequency body vibrations or near-field sound signals that differ from traditional far-field microphone signals, limiting the comparability between studies. Therefore, an integrative approach is required that takes advantage of the complementary strengths of the microphone and body-mounted sensor data to extract clean vocalization of each individual.

This challenge closely parallels Target Speech Extraction (TSE) in human speech processing. TSE aims to isolate the speech of a target speaker from a mixture of multiple speakers under noise and reverberation, using cues that identify the speaker in the mixture. Recent TSE systems exploit multiple cues [10], including audio cues from pre-recorded utterances of the target speaker [11, 12], spatial cues from microphone arrays [13, 14] or from video recordings of the overall scene [15], and visual cues such as lip movements captured through video analysis [16, 17]. Neural networks (NNs) typically integrate these diverse cues into robust pipelines for speech signal separation.

Following the TSE frameworks, we propose a vocalization separation system that uses body-mounted sensor signals as cues to separate overlapping vocalizations recorded with high-quality wall-mounted microphones. In this approach, accelerometer-based sensors mounted on animals record vocalizations via body vibrations, providing individual-specific cues. Additionally, multiple microphones provide spatially distinct audio recordings, potentially aiding in individual discrimination and sound separation. Building on an existing TSE framework [18], we propose Vib2Sound, a neural network for vocalization separation specifically designed to inte-

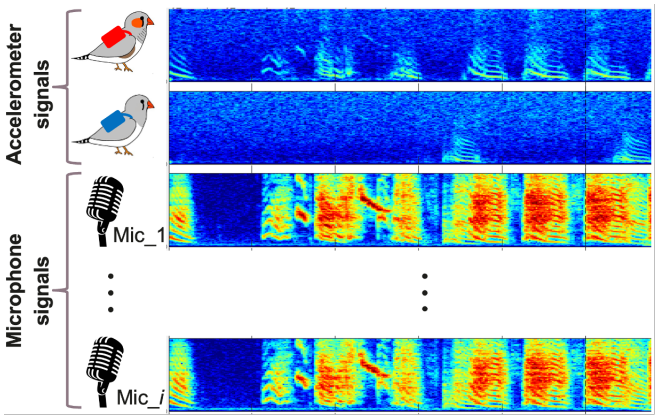

Fig. 1: Example synchronized data stream recorded in BirdPark [8]. The top two rows show spectrograms of body-mounted accelerometer signals, while the bottom spectrograms are derived from wall-mounted microphones.

grate individual-specific accelerometer cues and spatial information from microphone arrays.

We evaluate the efficacy of Vib2Sound on data collected in BirdPark [8], a sound isolated environment and system for synchronized multimodal recordings from animal groups. Each zebra finch in this setup carries an accelerometer mounted on its body, capturing the low-frequency vibration associated with vocalizations (Fig. 1). Specifically, our objectives are (1) to evaluate the performance of Vib2Sound in accurately separating overlapping vocalizations, (2) to investigate the extent to which body-mounted accelerometer cues enhance vocalization attribution compared to microphone-only methods, and (3) to assess the generalizability of Vib2Sound to realistic scenarios encountered in naturalistic settings. By addressing these objectives, our study aims to establish a robust methodological framework to advance individual-level analyses of vocal communication in animal groups, contributing valuable insights to research in behavioral ecology and bioacoustics.

2. METHODS

2.1. Proposed model

The structure of Vib2Sound is summarized in Fig. 2. Vib2Sound is based on VoiceFilter [18], but replaces the speaker-specific vocal embeddings with accelerometer-derived vibration cues and re-architects the model to operate on multi-channel audio. As inputs, Vib2Sound takes the mixed multi-channel microphone signals (Mic) and the respective accelerometer signals (Acc) from two birds. The network outputs two time-frequency masks that we used to predict the separated vocalizations of the two birds by applying them to the mixture spectrogram. The network is trained to minimize the mean squared error (MSE) between the predicted spectrograms and the two ground-truth spectrograms. We applied an inverse Short-Time Fourier Transform (iSTFT) to predict the demixed audio waveform.

The system consists of an 8-layer Convolutional Neural Network (CNN), 1 Long Short-Term Memory (LSTM) layer, and a 2-layer Fully Connected (FC) block. The magnitude (Mag) and phase spectrograms of the mixture signals from all channels are passed to the CNN to extract acoustic features (mixture audio stream). The magnitude spectrograms of the accelerometer signals (accelerometer stream) are concatenated to the CNN output and then passed to the LSTM layer. The output of the LSTM layer feeds into the FC layer to predict soft masks. The output soft masks are applied to the mixed magnitude spectrogram of the first microphone channel to estimate the target spectrograms, and the corresponding phase is used for waveform reconstruction.

2.2. Dataset

During training, the model requires clean isolated vocalizations. Following VoiceFilter, we artificially mixed microphone signals by summing two clean microphone audios.

Clean audios were extracted from recordings of 10 mixed-sex bird pairs in BirdPark, collected over 4 to 13 days per pair, with a sampling rate of 24414 Hz. We detected vocalizations using WhisperSeg [19], an automated vocal activity detection system. Non-overlapping vocalizations were manually selected as clean examples and divided into training, validation, and evaluation datasets, Table 2. For each bird, we reserved 10 % of the data as the evaluation dataset and 4.5 % as the validation dataset. We divided each dataset into two subsets based on the difference in microphone configuration, since spatial cues are highly sensitive to microphone position:

Dataset1 (14 birds): This setup includes 4 microphones. Mic1 and mic2 were centered on top of the right/front walls of the arena, mic3 on the back wall of the surrounding isolation box, and mic4 was installed on the ceiling of the box. The number of vocalizations per bird ranged from 430 to 1582 in the training dataset.

Dataset2 (6 birds): Mic1-4 are positioned above the 4 walls of the arena and mic5 on the ceiling of the isolation box. The number of vocalizations per bird ranged from 454 to 3263 in the training dataset.

	Training	Validation	Evaluation
Dataset1	8894	462	1033
Dataset2	8046	420	938

Table 1: Number of vocalizations in the training, validation, and evaluation datasets.

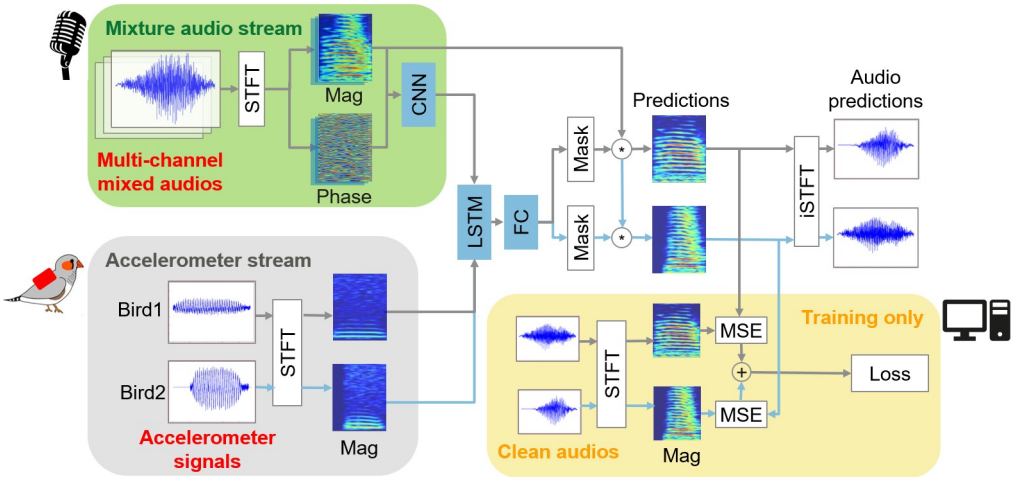

Fig. 2: Structure of Vib2Sound. Red indicates inputs to the model, and orange indicates inputs during training. In one training step, each accelerometer spectrogram is fed into the LSTM with the CNN output from the mixture audio stream, to compute the MSE loss for each bird (gray and light blue arrows). The sum of the 2 MSE losses is used as the training loss to be minimized.

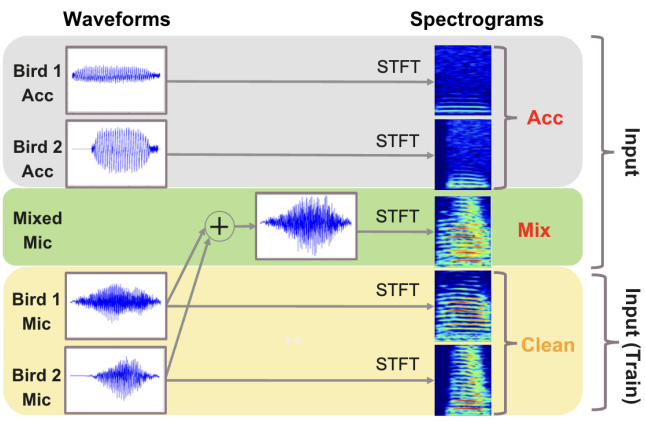

Fig. 3: Generating artificial sound mixtures for training.

2.3. Training

2.3.1. Synthetic mixture generation

We create synthetic mixtures by mixing pairs of randomly chosen clean vocalizations of two different birds in the training dataset, Fig. 3. To simulate realistic timing variability, we positioned the shorter audio segment at a random onset within the time span of the longer one.

2.3.2. Spectral features

We use a standard Short-Time Fourier Transform (STFT) to convert both the multi-channel microphone signals and the accelerometer signals into their complex (magnitude and phase) time–frequency representations. The STFT uses a fast Fourier Transform (FFT) length of 384 samples and a hop size of 96 samples. Additionally, we applied a 200 Hz high-pass filter to the accelerometer signal to remove low-frequency noise arising from radio signal transmission.

2.3.3. Training procedure

We trained one model for each of the two datasets (see Section 2.2). The batch size used for training was set to 4, and we used the Adam optimizer with a learning rate of 0.001. We trained the models for 200,000 steps.

2.4. Evaluation

2.4.1. Distance metrics

To evaluate the performance associated with diverse training strategies and models (see Section 2.4.2), we computed the distance score between the clean audio magnitude spectrogram and the predicted magnitude spectrogram on a time-bin basis. For each time bin t, we calculated the distance between the clean spectrum $\mathbf{S}_{\text{clean}}(t)$ and the predicted spectrum $\mathbf{S}_{\text{pred}}(t)$, and then we averaged the distances across all time bins of the vocalization:

$$\text{Distance Score} = \frac{1}{T}\sum_{t=1}^{T} \text{Dist}(\mathbf{S}_{\text{clean}}(t), \mathbf{S}_{\text{pred}}(t))$$

where T is the total number of time bins spanned by the vocalization, and Dist denotes the chosen distance metric (Euclidean, Cosine, or Spearman distance).

To assess performance, each model was evaluated on 1,000 artificial mixtures from the corresponding evaluation dataset. Each model generated 2,000 predicted vocalizations (two vocalizations from different birds per mixture), on which we evaluated distance scores.

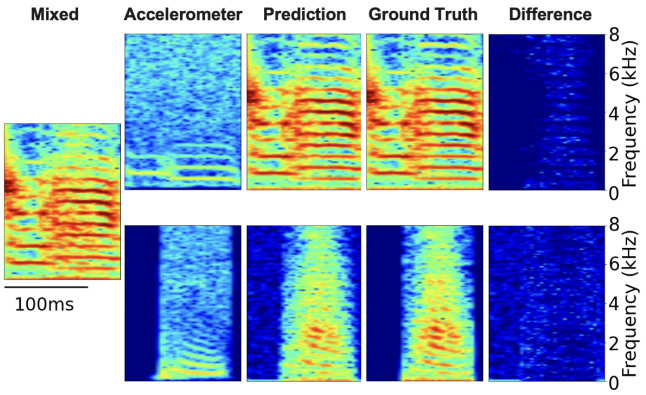

Fig. 4: Example of Vib2Sound's separation of an artificial mixture. From left to right, the spectrograms show: the artificially mixed microphone input, associated accelerometer signals, predicted vocalizations, ground-truth vocalizations, and absolute differences between the predicted and ground-truth spectrograms.

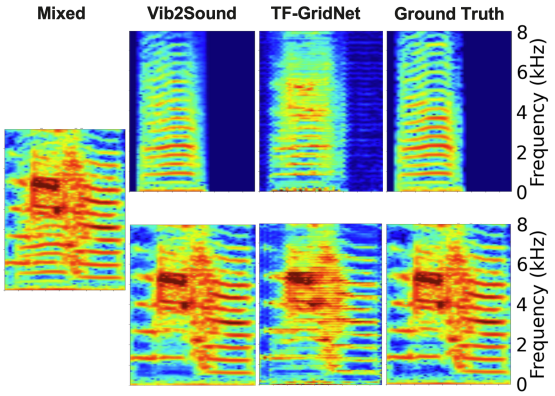

Fig. 5: Example prediction made by Vib2Sound and TF-GridNet.

2.4.2. Baseline models

We compared our proposed approach to the following baseline and ablation:

TF-GridNet [20]: TF-GridNet is a state-of-the-art speech separation model that leverages microphone array inputs. TF-GridNet operates in the complex spectrogram domain and involves two multi-channel-input single-channel-output (MISO) deep neural networks sandwiching a beamformer.

Single-channel Vib2Sound (ablation): To study the contribution of multi-channel microphone signals to the model, we constructed a single-channel Vib2Sound by feeding only the magnitude spectrogram from the first microphone into the CNN.

2.5. Real-world mixtures

In addition to evaluating the model on artificially mixed vocalizations, we also assessed its performance on naturally overlapping vocalizations extracted from our recordings. Since ground-truth clean sources are not available in these real-world cases, we conducted a qualitative visual inspection of the separation outputs. To further assess the plausibility of the predicted vocalizations, we employed a k-nearest neighbors (k-NN) retrieval approach to identify acoustically similar vocalizations from the clean training and validation sets. We used the `NearestNeighbors` implementation from the scikit-learn library, with Euclidean distance as the distance metric. This allowed us to evaluate whether the model's outputs were acoustically similar to real-world non-overlapping vocalizations, providing an additional qualitative check on the model's output consistency.

3. RESULTS AND DISCUSSION

3.1. Performance on artificial data

We summarize the experimental results, including distance scores, Multiply-Accumulate operations (MACs), and parameter sizes, in Table 2. The results highlight that Vib2Sound largely outperforms TF-GridNet across all distance metrics while being computationally more efficient.

To illustrate the separation performance qualitatively, Fig. 4 shows a representative example where Vib2Sound successfully separates overlapping vocalizations from two individuals. The predicted spectrograms closely match the ground-truth spectrograms with minimal residuals, indicating accurate reconstruction.

Our model outperformed TF-GridNet despite being more lightweight. Visual inspection reveals that although TF-GridNet correctly identified the overall vocalization boundaries, it failed to capture the fine-grained acoustic structure, Fig. 5. In contrast, Vib2Sound was able to reconstruct detailed vocalization patterns, leveraging the accelerometer signals.

The microphone array-based speech separation model did not perform well, likely due to several contributing factors. First, human speech, for which the model is built, is much more distinguishable among individuals with high variations in acoustic features such as pitch and formants, compared to zebra finch vocalizations. The short duration of bird vocalizations may also hinder the separation, as speech separation models typically rely on the contextual information obtained from a longer duration of speech. Additionally, the microphone configuration in the Birdpark differs significantly from the setups commonly used in human speech recordings, where spatial information is captured more effectively.

We further evaluated the role of multi-channel microphone input in the ablation experiment. As shown in Table 2, the single-channel model achieves nearly identical performance across all metrics. A t-test on the evaluation scores

	dataset1			dataset2			MACs (G)	params (M)
	Euclidean	Cosine	Spearman	Euclidean	Cosine	Spearman		
Mixture	1.668	0.0451	0.212	1.847	0.0575	0.215	-	-
multi-channel Vib2Sound	0.672	0.0257	0.149	0.537	0.0236	0.125	1.54	7.99
single-channel Vib2Sound	**0.665**	**0.0254**	**0.147**	**0.531**	**0.0229**	**0.124**	1.53	7.98
TF-GridNet	1.265	0.0511	0.268	1.239	0.0564	0.237	38.68	8.32

Table 2: Performance comparison of multi-channel, single-channel Vib2Sound and TF-GridNet. Bold values indicate the best performance. MACs and parameter sizes are computed for the 5-channel microphone inputs of 200 ms.

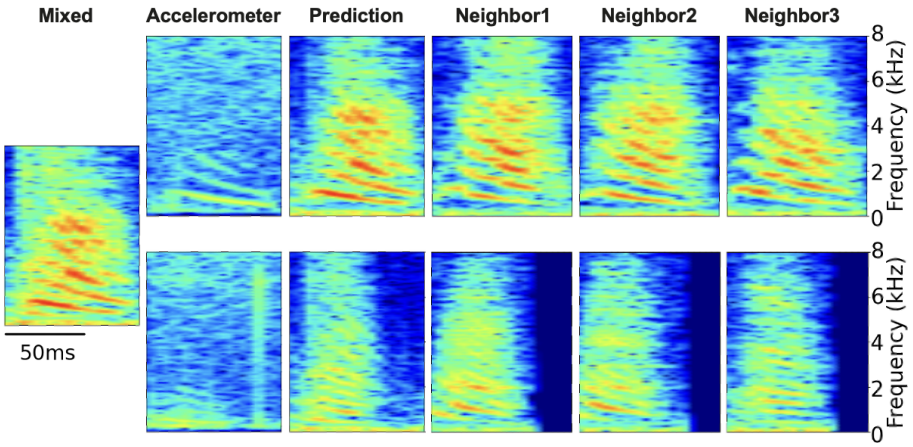

Fig. 6: Example of Vib2Sound prediction on naturally overlapping vocalizations. The right 3 columns show the top 3 nearest neighbors of the predicted vocalizations from clean training and validation datasets.

revealed no statistically significant differences caused by ablation, indicating that the model primarily relies on the accelerometer signals and a single mixed microphone channel for source separation.

3.2. Performance on real data

Fig. 6 shows that the vocalizations predicted by our model closely resemble clean vocalizations. The predicted spectrograms exhibit patterns that align well with those of the nearest neighbors, indicating that the model effectively captures the underlying structure of the data.

Fig. 6 shows an example in which no noise overlaps with the vocalizations. However, real-world data often contains various forms of noise arising from both animal behavior and environmental factors. To improve the model's robustness to such noise, future work should incorporate synthetic noisy data during training. This can be achieved by adding manually extracted noise to the mixed microphone and accelerometer signals, while retaining access to ground-truth clean audios for training loss computation.

4. CONCLUSION

In this work, we proposed Vib2Sound, a sound source separation model for animal vocalizations that works on microphone signals and leverages signals from body-mounted sensors as guiding cues. Evaluated on artificial mixtures of vocalizations of different individuals, our model outperformed a state-of-the-art speech separation model that relies on microphone-array inputs, highlighting the effectiveness of body-mounted sensor signals for identifying individual vocalizations during overlapping sounds. Furthermore, we demonstrated the reliability of our approach when applied to real-world overlapping vocalizations. This work facilitates the analysis of vocal communication in social animals, promoting deeper understanding of their social interactions and behaviors in natural settings.

5. REFERENCES

[1] Tessa A. Rhinehart, Lauren M. Chronister, Trieste Devlin, and Justin Kitzes, "Acoustic localization of terrestrial wildlife: Current practices and future opportunities," *Ecology and Evolution*, vol. 10, no. 13, pp. 6794–6818, 2020.

[2] Gongping Huang, Jesper R. Jensen, Jingdong Chen, Jacob Benesty, Mads G. Christensen, Akihiko Sugiyama, Gary Elko,

and Tomas Gaensler, "Advances in microphone array processing and multichannel speech enhancement," 2025.

[3] DeLiang Wang and Jitong Chen, "Supervised speech separation based on deep learning: An overview," 2018.

[4] Melinda L. Rekdahl, Rebecca A. Dunlop, Anne W. Goldizen, Ellen C. Garland, Nicoletta Biassoni, Patrick Miller, and Michael J. Noad, "Non-song social call bouts of migrating humpback whales," *The Journal of the Acoustical Society of America*, vol. 137, no. 6, pp. 3042–3053, 06 2015.

[5] Morgan L. Gustison, Aliza le Roux, and Thore J. Bergman, "Derived vocalizations of geladas (*Theropithecus gelada*) and the evolution of vocal complexity in primates," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 367, no. 1597, pp. 1847–1859, 2012.

[6] Shanah Rachel John, Rishika Tiwari, Yizhaq Goussha, Rotem Amar, Alex Bizer, Shai Netser, and Shlomo Wagner, "Simultaneous recording of ultrasonic vocalizations and sniffing from socially interacting individual rats using a miniature microphone," *Cell Reports Methods*, vol. 3, no. 11, pp. 100638, 2023.

[7] Elena N. Waidmann, Victor H.Y. Yang, William C. Doyle, and Erich D. Jarvis, "Mountable miniature microphones to identify and assign mouse ultrasonic vocalizations," *bioRxiv*, 2024.

[8] L. Rüttimann, J. Rychen, T. Tomka, H. Hörster, M. D. Rocha, and R.H.R. Hahnloser, "Multimodal system for recording individual-level behaviors in songbird groups," *bioRxiv*, p. 2022.09.23.509166, Jan. 2022.

[9] Zuyu Du, Yaodan Xu, Xinsheng Yu, Sen Wang, and Lin Xu, "Estimation of speech features using a wearable inertial sensor," *Journal of Voice*, 2024.

[10] Kateřina Žmolíková, Marc Delcroix, Tsubasa Ochiai, Keisuke Kinoshita, Jan Černocký, and Dong Yu, "Neural target speech extraction: An overview," *IEEE Signal Processing Magazine*, vol. 40, no. 3, pp. 8–29, May 2023.

[11] Alexey Ozerov and Emmanuel Vincent, "Using the FASST source separation toolbox for noise robust speech recognition," in *International Workshop on Machine Listening in Multisource Environments (CHiME 2011)*, Florence, Italy, Sept. 2011.

[12] Kateřina Žmolíková, Marc Delcroix, Keisuke Kinoshita, Tsubasa Ochiai, Tomohiro Nakatani, Lukas Burget, and Jan Cernocky, "Speakerbeam: Speaker aware neural network for target speaker extraction in speech mixtures," *IEEE Journal of Selected Topics in Signal Processing*, vol. PP, pp. 1–1, 06 2019.

[13] Rongzhi Gu, Lianwu Chen, Shi-Xiong Zhang, Jimeng Zheng, Yong Xu, Meng Yu, Dan Su, Yuexian Zou, and Dong Yu, "Neural Spatial Filter: Target Speaker Speech Separation Assisted with Directional Information," in *Proc. Interspeech 2019*, 2019, pp. 4290–4294.

[14] Jens Heitkaemper, Thomas Fehér, Michael Freitag, and Reinhold Haeb-Umbach, "A Study on Online Source Extraction in the Presence of Changing Speaker Positions," in *Statistical Language and Speech Processing*, Carlos Martín-Vide, Matthew Purver, and Senja Pollak, Eds., Cham, 2019, pp. 198–209, Springer International Publishing.

[15] Rongzhi Gu, Shi-Xiong Zhang, Yong Xu, Lianwu Chen, Yuexian Zou, and Dong Yu, "Multi-modal multi-channel target speech separation," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 3, pp. 530–541, Mar. 2020.

[16] John Hershey and Michael Casey, "Audio-visual sound separation via hidden markov models," in *Advances in Neural Information Processing Systems*, T. Dietterich, S. Becker, and Z. Ghahramani, Eds. 2001, vol. 14, MIT Press.

[17] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman, "The conversation: Deep audio-visual speech enhancement," *CoRR*, vol. abs/1804.04121, 2018.

[18] Quan Wang, Hannah Muckenhirn, Kevin Wilson, Prashant Sridhar, Zelin Wu, John Hershey, Rif A. Saurous, Ron J. Weiss, Ye Jia, and Ignacio Lopez Moreno, "Voicefilter: Targeted voice separation by speaker-conditioned spectrogram masking," 2019.

[19] Nianlong Gu, Kanghwi Lee, Maris Basha, Sumit Kumar Ram, Guanghao You, and Richard H. R. Hahnloser, "Positive transfer of the whisper speech transformer to human and animal voice activity detection," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 7505–7509.

[20] Zhong-Qiu Wang, Samuele Cornell, Shukjae Choi, Younglo Lee, Byeong-Yeol Kim, and Shinji Watanabe, "Tf-gridnet: Integrating full- and sub-band modeling for speech separation," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 31, pp. 3221–3236, Aug. 2023.